

A Dirty-Slate Approach to Routing Scalability

Hitesh Ballani, Paul Francis, Jia Wang and Tuan Cao
Cornell University and ATT-Research

Abstract—This paper presents *Virtual Aggregation*, an architecture that attempts to tackle the Internet routing scalability problem. Our approach does not require any changes to router software and routing protocols and can be deployed by any ISP without the cooperation of other ISPs. Hence, *Virtual Aggregation* is a configuration-only solution. The key insight here is to use divide-and-conquer so that default-free zone routers don't need to maintain the entire routing table. Instead, an ISP can modify its internal routing such that individual routers in its network only maintain a part of the routing table.

We evaluate the application of *Virtual Aggregation* to a few tier-1 and tier-2 ISPs and show that it can reduce routing table size on individual routers by an order of magnitude while imposing almost no traffic stretch and very little increase in router load. We also deploy *Virtual Aggregation* across two different testbeds comprising of Cisco and Linux routers. Finally, we detail some shortcomings of the proposed design and discuss alternative designs that alleviate some of these. However, in spite of the limitations, we believe that the simplicity of the proposal and its possible short-term impact on routing scalability suggest that it is an alternative worth considering.

I. INTRODUCTION

Today, the only means to control the size of the routing table in default-free zone (DFZ) routers is hierarchical aggregation of prefixes. For instance, use of Provider Aggregatable addresses by edge networks ensures that their prefixes can be aggregated by their providers. However, the rapid increase in the size of the DFZ routing table [22] suggests that hierarchy-based aggregation of prefixes has not been as effective we would desire. Contributors to this rapid growth of the routing table vary from technical factors such as multihoming by edge networks and traffic engineering to business events such as mergers and acquisitions [32]. Further, there are concerns that as the IPv4 address space runs out, aggregation will further deteriorate resulting in a substantial acceleration in the growth of the routing table [33]. Finally, a growing IPv6 deployment would worsen the situation even more [31].

The increase in the size of the DFZ routing information base (RIB) and forwarding information base (FIB) has several harmful implications for inter-domain routing in particular and the Internet in general. [33] discusses these in detail. At a technical level, increasing routing table size may drive high-end router design into various engineering limits. For instance, while memory and processing speeds might just scale with a growing routing system, power and heat dissipation capabilities may not [32]. A large routing table also causes routers to take longer to boot and exposes the core to edge dynamics, thereby afflicting routing convergence.

On the business side, increased memory requirements for both the RIB and the FIB and the performance requirements for forwarding while being able to access the FIB imply that the cost of forwarding packets increases and hence, networks become less cost-effective [29]. Further, it makes provisioning of networks harder since it is difficult to make estimates about the usable lifetime of routers, not to mention the cost of the actual upgrades. As a matter of fact, instead of upgrading their routers, a few ISPs have resorted to filtering out some small prefixes (mostly /24s) which implies that parts of the Internet don't have reachability to each other [20]. A recent conversation with a major Internet ISP revealed that in order to avoid router memory upgrades, the ISP is using a hack that reduces memory requirements but breaks BGP loop-detection and hence, would wreak havoc if adopted by other ISPs too. It is a combination of these possibilities that led a recent Internet Architecture Board workshop to conclude that scaling the routing system was one of the most critical challenges of near-term Internet design [32].

The severity of the routing scalability problem has also meant that a number of proposals have focussed on reducing the size of the DFZ routing table. One set of approaches try to reduce routing table size by dividing edge networks and ISPs into separate address spaces [6,10,31,34]. Alternatively, it is possible to encode location information into IP addresses [7,14,19] and hence, reduce routing table size. However, such location-based addresses place constraints on ISP inter-connectivity and do not allow the inter-connectivity to reflect ISP policies. Further, the aforementioned proposals require changes in the routing and addressing architecture of the Internet and perhaps this has contributed to the fact that none of them have seen deployment.

Guided by this observation, this paper takes an alternate approach towards the scalability problem and proposes *Virtual Aggregation*, an architecture that uses divide-and-conquer to engineer a scalable routing system. In our proposal, any given ISP can modify its internal routing such that individual routers in the ISP's network only maintain a part of the global routing table. This ensures that even though the growth of the DFZ routing table is not restricted, the demands placed on individual DFZ routers are. We argue that this alleviates most of the concerns arising out of the extreme growth in routing table size. To this effect, this paper makes the following contributions:

- We present a detailed *Virtual Aggregation* design that can be deployed independently and autonomously by any ISP. Further, the proposal applies to legacy routers and requires only configuration changes.

- We analyse the application of virtual aggregation to an actual tier-1 ISP and several inferred (Rocketfuel [40]) ISP topologies. We find that virtual aggregation can reduce routing table size by more than an order of magnitude with negligible average stretch on the ISP’s traffic and very little increase in load across the ISP’s routers. Based on predictions of future routing table growth, we estimate that virtual aggregation can be used to extend the life of already outdated routers by more than 10 years.
- We propose utilizing the notion of prefix popularity to reduce the impact of virtual aggregation on the ISP’s traffic and use a two-month study of a tier-1 ISP’s traffic to show the feasibility of such an approach.
- As a proof-of-concept, we configure two separate testbeds comprising of Linux software routers and Cisco routers (on WAIL [1]) according to the virtual aggregation architecture. We also configured virtual aggregation on the WAIL testbed using an alternative BGP-MPLS based configuration that reduces the management overhead of the deploying ISP.

Our proposal also suffers from a few drawbacks. While virtual aggregation can be achieved through appropriate configuration of existing routers, it does increase configuration complexity and impose management overhead. Also, the proposal does not reduce the total size of the global routing table per se. However, instead of offering a constant one-time improvement, it does improve the “scaling properties” of individual routers and hence, inter-domain routing. We discuss these and other shortcomings in section VI. However, in spite of these limitations, we believe that the simplicity of our proposal makes it an attractive short-term alternative that can help beyond measures being used today (such as FIB compression techniques, perilous hacks or even simply ignoring routes) and allow the routing system to cope with growing demands till more fundamental, long-term architectural changes can be agreed upon and deployed in the Internet.

II. WHY MOORE’S LAW WON’T SAVE US?

An oft-used argument regarding the increasing number of routable prefixes in the Internet is that Moore’s Law will ensure that both memory and processing power can scale with this growth. However, in a very interesting recent presentation [29], Tony Li argued against this by claiming that most components in a high-end router are not high volume entities and hence, do not follow the cost curve dictated by the Moore’s Law. For instance, both off-chip SRAM (used for storing the FIB) and ASIC processors used in high-end routers are past the inflection point in the cost vs performance curve. Consequently, while chip performance increases 2x every 2 years, chip costs are expected to grow 1.5x every 2 years. Hence, a growth in the routing system of more than 1.3x ($= 2/1.5$) every 2 years would make routing of packets less and less cost effective. And a look at past routing growth suggests that this has indeed been the case [21].

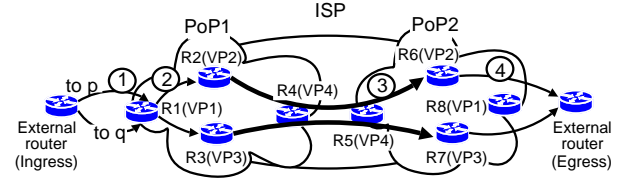


Fig. 1. An ISP with 4 virtual prefixes (VP1-VP4). Each router is an aggregation point for one virtual prefix; for ex, router R1 is an aggregation point for VP1. Prefix p belongs to VP2 and prefix q belongs to VP3. Traversal of packets through the ISP comprises of 4 parts.

Further, there are many router resources that Moore’s Law does not apply to. For instance, routers use DRAM to store the RIB and the fact that DRAM speeds grow only 1.2x every two years [28] affects the ability of routers to handle updates and hence, limits BGP convergence times. Finally and perhaps most importantly, the power consumption of high-end routers grows non-linearly with density increase. Thus, as routers require more memory and processing, their power consumption and the concomitant heat dissipation requirements could be a severe limiting factor for scaling [32].

While the arguments presented above certainly require more thorough investigation, they do suggest that the rapidly growing routing system is straining various router resources and hence, the need for reducing the routing load on routers. An obvious way to do this is to make the global routing table smaller, an approach used by past routing proposals. However, it is also possible to reduce the routing load on individual routers by ensuring that each router is only responsible for forwarding packets destined to a part of the address space. This, in turn, can be used to ensure that the memory, processing and heat dissipation requirements imposed on routers are closer to the point of inflection and hence, the cost of forwarding packets remains the same even as the global routing table grows.

III. ARCHITECTURE

The key insight behind the *Virtual Aggregation* architecture is to allow individual ISPs in the Internet’s DFZ to do away with the need for their routers to maintain routes for all prefixes in the global routing table. Instead, only the ISP’s route-reflectors that are not in the data-path and hence, do not forward packets need to maintain the entire routing table.¹ To this effect, an ISP desiring to reduce routing load on its routers divides the global address space into a set of *virtual prefixes*. For instance, an ISP could divide the IPv4 address space into 128 parts with a /7 representing each part (0.0.0.0/7 to 254.0.0.0/7). Note that these /7 prefixes are not topologically valid aggregates, i.e. there is not a single point in the Internet topology that can hierarchically aggregate the encompassed actual prefixes and hence, the term *virtual prefixes*.

With such a division in place, the ISP can modify its internal routing so that each router in the ISP’s network

¹Later in the paper, we explain why this is reasonable and how the route-reflectors can be scaled.

only maintains routes for prefixes in one (or, a few) virtual prefix.² A router that maintains routes for prefixes in a given virtual prefix is an *aggregation point* for the virtual prefix. For example, figure 1 illustrates an ISP using 4 virtual prefixes with an aggregation point for each virtual prefix in each PoP. The aggregation points for a given virtual prefix are then organised into a tunneled topology that is the *virtual network* for the virtual prefix. In effect, the virtual networks allow for efficient aggregation of the virtual prefixes. In figure 1, router R2 and R6 are aggregation points for virtual prefix VP2 and hence, are connected by a tunnel. The figure also shows that a typical path through the network for packets destined to a prefix p belonging to VP2 comprises of 4 parts: (1). A native path from the external router to edge router R1, (2). A native path from R1 to aggregation point R2, (3). A tunneled path from R2 to R6 and, (4). A native path from aggregation point R6 to the external router at the egress.

The discussion above describes the operation of virtual aggregation at a conceptual level. However, the design space for the actual deployment of virtual prefixes in an ISP's network is characterized by several dimensions. For example, the flexibility to add devices to the ISP's network, to change the ISP's topology or to change the routers themselves lead to very different architectures, all of which allow for virtual prefix based routing. While we discuss some such alternative approaches in section VI, this paper focusses on one particular design guided by two major design goals:

- (a). *No changes to router software and routing protocols*: The ISP should not need to deploy new data-plane or control-plane mechanisms.
- (b). *Transparent to external networks*: An ISP's decision to adopt the virtual aggregation proposal should not impact its interaction with its neighbors (customers, peers and providers). For instance, on the control-plane side, the ISP's eBGP peerings with external routers should not be affected.

A. Overview

The key challenge in virtual aggregation is to ensure that all four parts of the path shown in figure 1 can work while satisfying the aforementioned design goals:

- Segment (1) involves packets being routed to the ISP's edge router and hence, does not require any special mechanisms.
- For segment (2), packets from edge routers need to be directed to the ingress aggregation point. We achieve this by ensuring that each router knows about an aggregation point for each virtual prefix.
- For segment (3), packets need to be tunneled through the virtual network. All routers used in ISP networks today

support many tunneling protocols, including IP-IP, GRE-IP, MPLS, etc.

- For segment (4), packets from the egress aggregation point need to get to the external router without any "routing" in the middle. We achieve this by using the BGP *next-hop* attribute, ensuring that the egress aggregation point and the external router have layer-2 connectivity between them and using ARP to steer packets to the appropriate external router.

The following sections detail these mechanisms and describe how an ISP can deploy virtual aggregation. The discussion below applies to IPv4 (and BGPv4) although the techniques detailed here work equally well for IPv6.

All the ISP's routers participate in an intra-domain routing protocol that establishes internal routes through which the routers can reach other. For each virtual prefix, the ISP designates some number of routers to serve as aggregation points for the prefix. For ease of exposition, the discussion in the following sections assumes that each PoP in the ISP has at least one aggregation point for each virtual prefix. We discuss how this condition can be relaxed in section III-G. As in networks today, the ISP establishes eBGP peerings with (external) routers belonging to neighboring ASes and obtains all routes advertised by the neighbor. However, these routes cannot reside on a single router. Instead, routes for prefixes that belong to a given virtual prefix need to be sent to the aggregation point(s) for the virtual prefix in the PoP that the external router is connected to. This process is detailed in section III-B. Beyond this, the aggregation points for a virtual prefix in different PoPs exchange routing information so that each aggregation point has routes for all prefixes in the corresponding virtual prefix (section III-C). Finally, routers that are not aggregation points for a virtual prefix should be able to send packets destined to prefixes in the virtual prefix and we discuss the mechanism used to achieve this in section III-D.

B. External peerings

Control Plane: The ISP uses eBGP peerings to exchange routing information with its neighbors. However, the edge routers for the ISP cannot establish these peerings since that would entail the routers needing to keep all the routes being advertised by the neighbors in their RIB and FIB and thus, maintaining the full DFZ routing table.³ Consequently, we offload the task of interacting with the external networks connected to each of the ISP's PoPs to a separate entity in the PoP that should satisfy the following condition:

It should not be in the data path so that the size of the FIB is not of critical concern.

This entity effectively serves as a conduit for the exchange of routes between the PoP's internal and external

²Prefixes that are more specific than a virtual prefix and hence, are encompassed by it are referred to as being "in the virtual prefix".

³ISP networks typically comprise of a few kind of routers: *core/backbone* routers, *aggregation* routers, *peering/edge* routers and *access* routers. In the rest of this paper, we abuse terminology and refer to all non-core routers as *edge* routers.

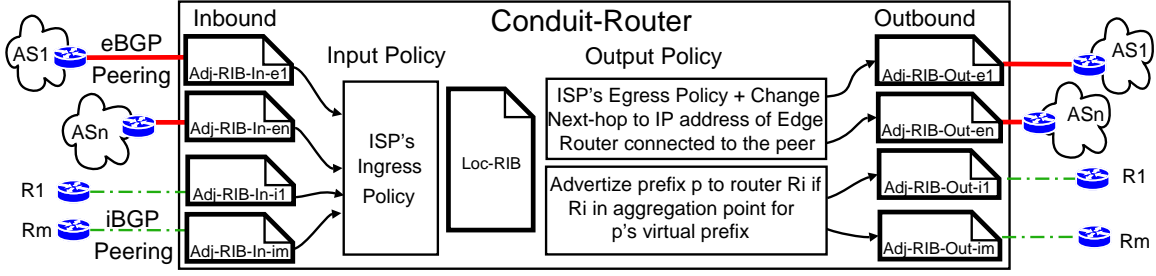


Fig. 2. The conduit-router for a PoP is a vanilla router that uses ingress-egress filters to ensure that it is off the data path while appropriately exchanging routes between the PoP's internal and external routers. All this configuration can be done on *existing* routers. Here, router R1 to Rm are the PoP's internal routers while AS1 to ASn are the ISP's neighbors connected to the PoP.

routers and hence, is referred to as the *conduit-router* for the PoP. Note that the conduit-router does need to maintain the full DFZ routing table and we describe later how its RIB can be scaled. Further, the conduit-router can be and will be replicated in each PoP. Figure 2 depicts a conceptual view of the operation of a conduit-router. The figure shows that, despite the name, the conduit-router for a PoP is simply a vanilla BGP router that establishes eBGP peerings with all the external routers connected to the PoP and iBGP peerings with all the PoP's internal routers. The conduit-router installs the routes that it receives from all its peerings into its corresponding adjacency RIB (Adj-RIB-In) and then uses the BGP decision process to determine its local-RIB [48].

As far as outbound route advertisements to its iBGP peers are concerned, the conduit-router needs to advertise routes to the internal routers such that each router only receives routes for prefixes in the virtual prefix it is aggregating. As shown in figure 2, such demultiplexing of routes can be achieved through the use of egress-filters on individual iBGP peerings that appropriately restrict the routes advertised to the peer. We discuss the design of such filter rules in section V. Note that through these advertisements, the conduit-router is effectively redistributing routes obtained from eBGP peers to its iBGP peers. The default BGP behavior in such a scenario is the *next-hop* attribute in the corresponding advertisements to be set to the IP address of the eBGP peer the route was originally obtained from [51]. Hence, as long as the conduit-router is not on the physical path between the internal and external routers, it is not on the data-path of packets flowing out of the ISP's network.

The conduit-router, based on the ISP's policy, also advertises routes in its loc-RIB to its eBGP peers. However, in order to satisfy the aforementioned off-path condition for packets flowing into the ISP's network (segment (1)), the conduit-router changes the *next-hop* attribute in the routes to the IP address of the edge router that the eBGP peer is physically connected to. As shown in figure 2, this is achieved through egress-filters for the eBGP peerings that modify the *next-hop* of the route being advertised appropriately.

Data Plane: The aforementioned control-plane mechanisms essentially establish segments (1) and (4) in figure 1.

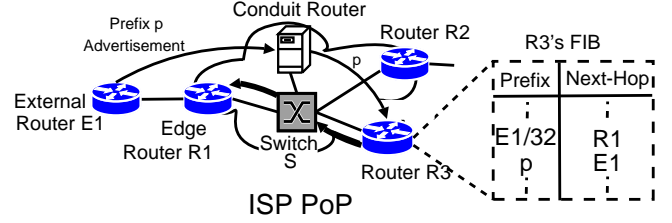


Fig. 3. Packets destined to prefix p are forwarded by router R3 to next-hop E1. However, edge router R1 cannot forward these packets since it does not have a route for p in its FIB.

Segment (1) is the same as in today's set-up. Segment (4) is established by routes obtained from external routers that are advertised to the respective aggregation points in the PoP with the *next-hop* attribute of the route set to the external router originating the route. For such a route to be usable, the external router should be reachable from the ISP's routers. This is accomplished today by advertising a route to the external router into the ISP's IGP. However, with virtual aggregation, the fact that only a fraction of the ISP's routers have a FIB entry to a given prefix implies that such an approach does not work. This problem is illustrated in figure 3. In the figure, external router E1 advertises a prefix p to the conduit-router. Also, it is assumed that router R3 is an aggregation point for p's virtual prefix and hence, the conduit-router advertises prefix p to router R3. When R3 receives a packet destined to prefix p, it forwards it onto next-hop E1. However, once these packets reach router R1, they cannot be forwarded since R1 is not an aggregation point for p's virtual prefix and hence, does not contain a route for p. This problem could be avoided by tunneling packets from router R3 to E1 but that would require cooperation from the neighboring ISP.

The solution to this problem is to ensure layer-2 connectivity between routers E1 and R3 so that none of routers along the path between them need to "route" the packets. More generally, such an approach implies that there needs to be layer-2 connectivity between all the routers of a PoP and the external routers that the PoP peers with. Today, ISP PoPs generally comprise of an "access tier" made up of edge and peering routers, an "aggregation tier" made up of layer-2 switches and a "core tier" made up of core routers [38]. Thus, as illustrated in figure 3, all of the PoP's routers are already connected at layer-2. Further, almost all

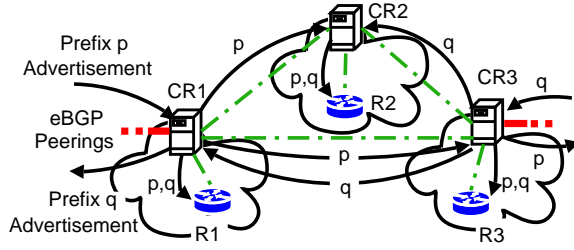


Fig. 4. Conduit-router in a PoP peers with other conduit-routers and reflects routes onto the appropriate aggregation points in its PoP. Prefixes p and q belong to the same virtual prefix and the figure only shows the aggregation points for this virtual prefix in each PoP.

routers today, including the ones that we used as part of our deployment effort, have both layer-2 switching and layer-3 routing capabilities.

Given this, the PoP's edge routers such as R1 are configured to also switch packets at layer-2. Consequently, if the PoP-facing interface of an external router (i.e. E1) is assigned an address that is on the same subnet as the rest of the PoP's routers, the PoP's routers will have layer-2 connectivity to the external router. With such an arrangement, when router R3 forwards a packet destined to prefix p, it notices that next-hop router E1 is on the same subnet, uses an ARP lookup to determine the MAC address for E1 and forwards the packet which is switched at layer-2 by S and R1 onto E1. Note that peerings between ISPs anyway involve co-ordination of the addresses and subnet masks to be used for the peering and hence, from a technical perspective, the fact that the external router needs to be on the same subnet as the PoP's routers does not impose any additional burden on the ISP's neighbors.

Such an arrangement also implies that the conduit-router has layer-2 connectivity to external routers (see figure 3). Hence, eBGP peerings between them involve a single IP hop. This is important since ISPs are generally averse to establish multihop eBGP peerings [44].

C. Virtual networks for virtual prefixes

Control-Plane: The aggregation points for a virtual prefix in the ISP's PoPs need to exchange routes among themselves for prefixes obtained from external routers via the conduit-router of their PoP. Note that this task is similar to the distribution of external routes in the ISP's network with today's setup. The trivial way of achieving this is to establish a complete mesh of iBGP peerings between the ISP's routers [45]. Such a mesh of iBGP peerings raises obvious scalability concerns and hence, ISPs commonly use route reflectors [45] and confederations [36] as a scalable means of distributing external routes.

In virtual aggregation, the conduit-router of each PoP already has peerings with the PoP's routers and hence, can also be used as a route-reflector for the PoP. Consequently, as shown in figure 4, the conduit-routers in different PoPs have a mesh of iBGP peerings with each other, though they

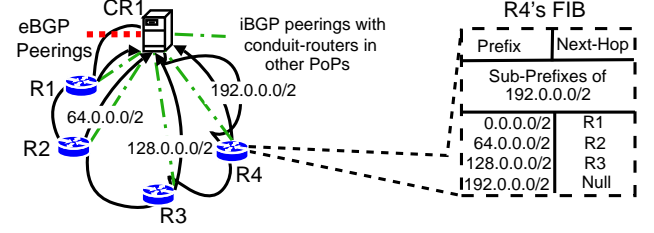


Fig. 5. Routers of a PoP exchange virtual prefixes through the PoP's conduit-router. The ISP is using 4 virtual prefixes (/2s) with R1 aggregating 0.0.0.0/2, R2 aggregating 64.0.0.0/2, R3 aggregating 128.0.0.0/2 and R4 aggregating 192.0.0.0/2.

could just as well be arranged in a multi-level hierarchy. The figure also shows the propagation of advertisements for a couple of prefixes through the ISP's network. Note that in practice, the ISP would just use its existing route-reflectors to operate as conduit-routers.

Data Plane: Using the conduit-routers as route-reflectors allows for an exchange of routes between aggregation points for a virtual prefix in different PoPs and hence, establishes segment (3) in figure 1. However, these aggregation points are not directly connected to each other and the routers along the path may not have routing information for the virtual prefix in question. This implies that the aggregation points need to tunnel packets between each other. Consequently, the aggregation points are connected using a set of tunnels and this forms the *virtual network* associated with the virtual prefix. The virtual network may be formed in a number of ways, as long as it is connected. For instance, it may be a full mesh with a tunnel between every pair of aggregation points. Or, tunnels may be established only between aggregation points in PoPs that share a physical link. An intra-domain routing protocol can then be used over this virtual network to ensure that all aggregation points have tunneled reachability to each other.

The use of a virtual network also has implications for the peerings between the conduit-routers. When a conduit-router advertises a prefix to other conduit-routers, it modifies the next-hop attribute in the advertised routes to the ip-address of the tunnel interface of the router in its PoP that aggregates the corresponding prefix. For instance, in figure 4, the next-hop in prefix p's advertisement to CR2 and CR3 contains R1's tunnel interface. This ensures that packets forwarded between the aggregation points are tunneled and hence, use the virtual network in place.

D. Connecting virtual networks

Control Plane: A router has routes for all prefixes in the virtual prefixes it is aggregating and hence, can forward packets destined to such prefixes appropriately. For any other prefix, the router needs to forward packets destined to the prefix to the nearest router that is an aggregation point for the virtual prefix encompassing the prefix. In other words, routers that are not aggregation points for a virtual prefix need to know how to get to the virtual

network associated with the virtual prefix. This corresponds to segment (2) in figure 1

To achieve this, each router in a PoP *originates* a route for the virtual prefixes it is aggregating. This advertisement is propagated by the PoP's conduit-router to other routers in the PoP. Figure 5 illustrates this using a 4-router PoP in an ISP that is using 4 virtual prefixes. Router R4 receives routes for the 3 other virtual prefixes and hence, can reach the corresponding virtual networks. Note that the routers only exchange virtual prefixes and not any of the specific prefixes. Hence, each router that is not an aggregation point for a virtual prefix only needs to maintain one FIB entry for the virtual prefix.

E. Conduit-router scalability

The fact that the conduit-router is not on the data path implies that its FIB needn't be on fast memory and hence, the FIB size is not critical. However, the conduit-router for a PoP peers with all the external routers connected to the PoP. The RIB size on a BGP router depends on the number of peers it has and hence, the RIB for the conduit-router can potentially be very large. However, we can scale the RIB requirements by using a hierarchy of machines to peer with external routers and feed into the PoP's conduit-router. Note that while these machines and the conduit-router still need to maintain the full DFZ routing table, the resulting RIB scaling properties are better than in the status quo. Today, edge routers have no choice but to peer with the directly connected external routers and maintain the resulting RIB. Replicating these routers is prohibitive because of their cost but the same does not apply to our proposal. The fact that none of the machines peering with external routers are on the data path implies that they can even be BGP software routers running on PCs.

F. Network robustness

The use of virtual aggregation by an ISP raises many issues, none more important than its impact on the robustness of the ISP's data and control plane. On the data-plane side, the use of virtual aggregation implies that a packet traversing the ISP's network needs to go through the aggregation point for the destination prefix in the ingress and the egress PoP. Further, there is the issue of packets being tunneled between aggregation points and the concomitant robustness concerns. While we address the tunnel maintenance issues in section V, the ISP can avoid a single point of failure for a virtual prefix's traffic in each PoP by ensuring that more than one router per PoP aggregates the virtual prefix. Thus, by controlling the *replication factor* (RF), the ISP can tune data-plane robustness and ensure that it is qualitatively no worse than today. On the control-plane side, the conduit-router in virtual aggregation is the same as the per-PoP route-reflector used by ISPs today and will be replicated for robustness.

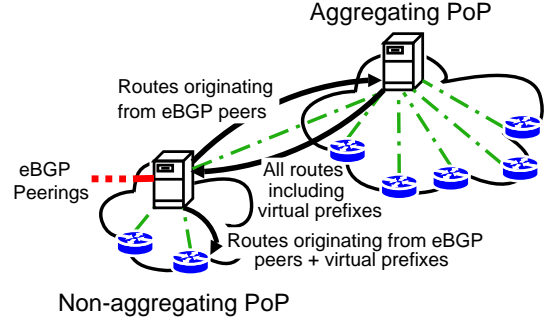


Fig. 6. Conduit router of a non-aggregating PoP is a route-reflector client for the conduit-router of a neighboring aggregating PoP.

G. Aggregating PoPs

The sections above assume that each PoP has at least one aggregation point for each virtual prefix. If the ISP is using V virtual prefixes, the number of virtual prefixes per router in a PoP with N routers is $(V * RF)/N$. Also, assuming an even distribution of prefixes across virtual prefixes, the size of the global routing table is $V * (\text{virtual prefix size})$.⁴ Hence, the FIB size for a router in the PoP is given by,

$$\begin{aligned} \text{FIB Size} &\approx (\text{Virtual prefix size}) * (V * RF)/N \\ &\approx \frac{(\text{Global Routing Table Size})}{N} * RF \end{aligned}$$

Thus, the size of the FIB is (roughly) inversely proportional to the number of routers in the PoP. However, such a deployment implies that the advantages of virtual aggregation would be severely limited by small PoPs. This problem applies not only to small tier-2 and tier-3 ISPs but also to tier-1 ISPs where PoP sizes show a significant variation. For instance, an analysis of the Rocketfuel topologies [40] of 10 tier-1 and tier-2 ISPs shows that 6 ISPs have at least one PoP of size 2 while 3 have at least one PoP of size 3. Note that a PoP with two routers in an ISP using a replication factor of two implies that the routers of the PoP need to maintain the full DFZ routing table and this defeats the purpose of using virtual aggregation.

Hence, when an ISP deploys virtual aggregation, it should be able to choose which of its PoPs will have aggregation points for virtual prefixes (*aggregating PoPs*) and which will not (*non-aggregating PoPs*). The key idea behind the operation of a non-aggregating PoP is that while it can forward traffic that is to be routed through the PoP's external routers on its own, it relies on a close-by aggregating PoP to forward the rest of its traffic. We briefly explain this below.

Control-Plane: Figure 6 illustrates the operation of a non-aggregating PoP. As before, the non-aggregating PoP has a conduit-router that serves as a route-reflector for the PoP's internal routers and peers with the PoP's external routers. However, it also serves as the route-reflector client of the conduit-router of at least one aggregating PoP. Hence, the conduit-router gets the routes for all virtual prefixes and

⁴Virtual prefix size refers to the number of prefixes in the virtual prefix. Also, we justify the assumption in section VI.

all prefixes inside the virtual prefixes. The conduit-router advertises the virtual prefixes to the PoP's internal routers. However, the conduit-router only advertises a route to a non-virtual prefix to the internal routers if the route was obtained the one of its eBGP peers. This ensures that routers in non-aggregating PoPs have FIB entries for all prefixes that need to be routed through external routers connected to the PoP while for all other prefixes, they rely on the aggregating PoP that their conduit-router peers with.

Data-Plane: The conduit-router of a non-aggregating PoP advertises routes obtained from its eBGP peers to the conduit-router of the aggregating PoP it peers with. Hence, the rest of the network receives routes from neighboring ASes that are connected to the non-aggregating PoP. However, for reasons discussed in section III-C, the actual data packets from the routers of aggregating PoPs to the non-aggregating PoP need to be tunneled. Consequently, each router of the non-aggregating PoP needs to have tunneled reachability to the aggregation points of all virtual prefixes and hence, is a member of the virtual network associated with each virtual prefix.

There are a couple of other ways to allow for non-aggregating PoPs. Further, the discussion above presents a very simple, coarse-grained approach wherein an aggregating PoP has aggregation points for all virtual prefixes. This approach can be extended to ensure that a given PoP only aggregates some of the virtual prefixes while relying on neighboring PoPs to route packets destined to prefixes in other virtual prefixes. This would be useful for smaller ISPs wherein all the PoPs have a few routers. While non-trivial, in the interest of brevity we don't discuss these extensions in this paper although we do come back to the issue of tier-2 and tier-3 ISPs later in the paper.

The discussion above suggests that an important aspect of virtual aggregation deployment by an ISP is to choose which of its PoPs should be aggregating PoPs. Note that this choice represents a trade-off between the size of the FIB on the routers, the stretch imposed on traffic and router load since non-aggregating PoPs rely on a nearby aggregating PoP for routing traffic to most of the prefixes in the DFZ routing table. Hence, the ISP would need to strike a balance between these factors. From a practical perspective, many ISPs today already have a tiered structure with satellite PoPs in small cities feeding traffic into a few major PoPs [16]. Such tiering of PoPs fits naturally into the choice of aggregating and non-aggregating PoPs. We examine this choice of aggregating PoPs and the aforementioned trade-off for an Internet tier-1 ISP in section IV-A.

H. Routing popular prefixes natively

Virtual aggregation causes packets to take paths longer than native paths. When packets traverse an aggregating PoP, they need to be routed through the aggregation point for the destination prefix. Since the extra links traversed are intra-PoP links, the actual increase in path length is minimal. However, the same cannot be said for non-aggregating

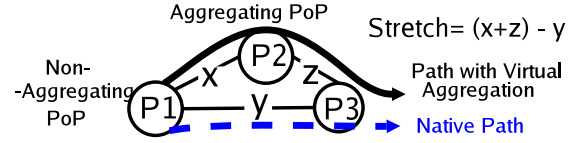


Fig. 7. Packets from non-aggregating PoPs to some prefixes may take paths longer than native paths.

PoPs wherein packets need to traverse extra inter-PoP links as part of being backhauled to an aggregating PoP. The packets also incur queuing delay at all the extra hops. Finally and perhaps most importantly, the extra hops impose extra load on the routers and modify the distribution of traffic across the routers.

On the other hand, past studies have shown that a large majority of Internet traffic is destined to a very small fraction of prefixes [9,12,37,46]. The fact that routers today have no choice but to maintain the complete DFZ routing table implies that this observation wasn't very useful for routing configuration. However, with virtual aggregation, individual routers only need to maintain routes for a fraction of prefixes. The ISP can thus configure its virtual aggregation setup such that the small fraction of popular prefixes are in the FIB of every router and hence, are routed natively. Such dissemination of popular prefixes can be easily achieved by ensuring that conduit-routers don't filter out the advertisements for the popular prefixes to any of the internal routers. The rest of the proposal involving virtual prefixes remains the same and ensures that individual routers only maintain routes for a fraction of the unpopular prefixes. In section IV-B, we analyze Netflow data from a tier-1 ISP network to show that not only such an approach is feasible, it also addresses all the concerns raised above.

IV. EVALUATION

In this section we evaluate the application of virtual aggregation to a few Internet ISPs. The main results presented here are:

- Using data from a tier-1 ISP we show that virtual aggregation can reduce the FIB size by a factor of more than 10 with negligible stretch on the ISP's traffic and a very small, gradual increase in router load. Given predictions of future routing table growth, we find that virtual aggregation would allow ISPs to extend the life of outdated routers by more than 10 years.
- Based on a two-month long study of the ISP's traffic we conclude that prefix popularity can indeed be used to minimise the impact of virtual aggregation on the ISP's traffic.
- We analyse the application of virtual aggregation to the Rocketfuel topologies of 10 ISPs and conservative estimates show that in the worst case, the FIB size on the ISP's routers is reduced to less than 15% of the DFZ routing table.

A. Tier-1 ISP Study

We simulated the application of virtual aggregation to a large tier-1 ISP in the Internet. For our study, we obtained

the ISP's router-level topology and using the location of the routers, mapped it to the ISP's PoP-level topology. Further, we annotated the inter-PoP links in this topology with latency information based on the geographical locations of the PoPs. We also obtained the BGP routing tables used by the ISP's routers and the ISP's PoP-level traffic matrix.

To apply virtual aggregation to the ISP's network, we divide the IPv4 address space into 128 parts and use /7s as virtual prefixes. The fact that the ISP has a few small PoPs implies that we need to designate some PoPs as non-aggregating PoPs. For robustness, routers in aggregating PoPs aggregate virtual prefixes such that each virtual prefix has two aggregation points ($RF=2$). Further, routes to 1.5% of the most popular prefixes are maintained by all ISP routers.⁵ We show in section IV-B that, on average, these prefixes carry 75.5% of the ISP's traffic. Given this, the discussion below focusses on the following parameters:

- **FIB Size**: Apart from routes to popular prefixes, routers in aggregating PoPs only maintain routes to prefixes in the virtual prefixes they are aggregating and routes to the virtual prefixes themselves. Routers in non-aggregating PoPs only maintain routes to prefixes that need to be routed out of the ISP's network through them and routes to the virtual prefixes and popular prefixes. This is typically a small number of routes and hence, we focus on the FIB size in routers of aggregating PoPs. We define the **average FIB size** as the average fraction of the DFZ routing table that the ISP's routers need to maintain. However, the more interesting metric is the maximum number of FIB entries that any of the ISP's routers need to maintain. As described in section III-G, the size of the FIB that a router in an aggregating PoP needs to maintain depends on the number of routers in the PoP. Hence, the **worst-case FIB size** applies to routers in the smallest ISP PoP that is designated as an aggregating PoP.

- **Stretch**: Non-aggregating PoPs rely on a nearby aggregating PoP to route traffic to most destinations. In figure 7, non-aggregating PoP P1 peers with aggregating PoP P2. Hence, packets to unpopular prefixes will be routed through PoP P2 even though the native (and shorter) path for some of these prefixes may directly go to PoP P3. Using today's routing table for individual routers and information regarding the latency between the ISP's PoPs, we can calculate the stretch that virtual aggregation would impose on any non-aggregating PoP's traffic to any Internet prefix. To characterize the most unfavorable impact of the use of virtual aggregation, we look at **worst-case stretch** which is the maximum stretch imposed on any prefix's traffic across all the ISP's PoPs.

However, worst-case stretch does not depict the entire picture since it does not account for the amount of traffic that suffers the stretch. Hence, we define the **traffic-averaged stretch** (or, simply **average stretch**) as the

⁵In a practical deployment, the ISP will also put routes to its managed customers in all routers so that the corresponding traffic uses native paths.

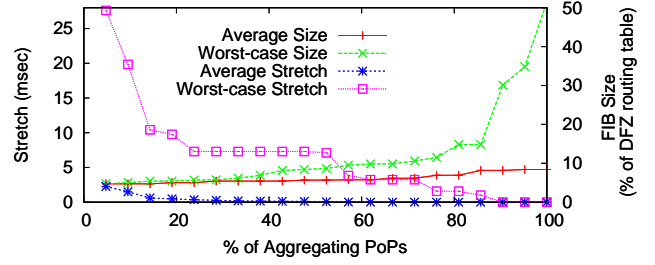


Fig. 8. Variation of traffic stretch and FIB size as the ISP uses more of its PoPs to aggregate prefixes.

amount of stretch averaged across all the ISP's traffic. Specifically,

$$\text{AverageStretch} = \sum_{\substack{i \text{ is non-agg} \\ \text{regating PoP}}} \text{max-stretch}_i * \text{unpopular-traffic}_i$$

where, max-stretch_i is the maximum stretch imposed on PoP i 's traffic to any prefix and $\text{unpopular-traffic}_i$ is the amount of traffic to unpopular prefixes from PoP i as a fraction of the ISP's total traffic. Note that both these metrics only account for delay due to increased distance, not the queuing delay imposed by the additional routers traversed.

- **Traffic**: We define **traffic impacted** as the fraction of the ISP's traffic that uses a different router-level path than the native path. This, in turn, imposes extra load on the routers. The **load increase** across a router is the extra traffic it needs to forward due to virtual aggregation, as a fraction of the traffic it forwards natively. The extra load is especially critical for edge routers that have relatively low bandwidth interfaces and hence, we measure the **average load increase** across the ISP's edge routers. Another metric of interest is the **traffic stretched**, the fraction of traffic that is forwarded along a different PoP-level path than before. In effect, this represents the change in the distribution of traffic across the ISP's inter-PoP links and hence, captures how virtual aggregation interferes with the ISP's inter-PoP traffic engineering. Note that only traffic from non-aggregating PoPs can use a PoP-level path different than before and hence, would contribute to this metric.

In order to minimise the FIB size on its routers, the ISP would like to use large PoPs as aggregating PoPs. To study the impact of such an allocation, we sort the ISP's PoPs based on their size and designate the top-k as aggregating PoPs; i.e., when the ISP needs to choose four aggregating PoPs, it chooses the four largest PoPs. This represents a greedy assignment strategy. The Y-axis on the right in figure 8 shows how the FIB size varies with the percentage of the ISP's PoPs that are designated as aggregating PoPs. As expected, both the average and the worst-case FIB size increase as more (and hence, smaller) PoPs start getting designated as aggregating PoPs. The figure shows that till $2/3^{rds}$ of the PoPs are designated as aggregating PoPs, the average FIB size stays less than 6% and the worst-case FIB size stays less than 10% of the DFZ routing table. On the other hand, both the average and the worst-case

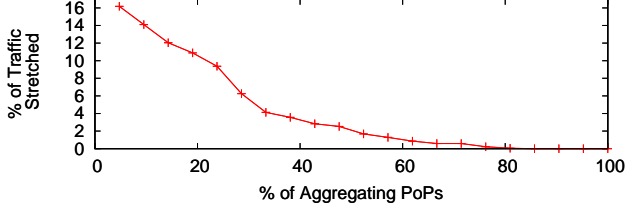


Fig. 9. Variation of the traffic stretched as the percentage of aggregating PoPs increases.

% of Ag. PoPs	FIB Size (%)		Stretch (msec)		Traffic (%)	
	Average	Worst	Average	Worst	Impacted	Stretched
57	5.75	9.5	0.023	3.81	24.5	1.3
86	8.17	14.83	~0	0.99	24.5	~0

TABLE I

DEPLOYMENT PARAMETERS WITH WORST-CASE STRETCH
CONSTRAINED TO 5MSEC (LINE 1) AND 1MSEC (LINE 2)

stretch drop sharply as the percentage of aggregating PoPs increases. This is because not only does the number of non-aggregating PoPs reduce, but for each non-aggregating PoP, there is a greater chance of a nearby aggregating PoP and hence, the stretch imposed on its traffic reduces too. When more than half the PoPs are serving as aggregating PoPs, the average stretch imposed is less than 0.06 msec and the worst-case stretch is less than 4 msec. The average stretch is almost negligible due to two reasons. First, the smaller PoPs are designated as non-aggregating PoPs and these generate a relatively smaller fraction of the ISP's total traffic. Second, only traffic to the unpopular prefixes from non-aggregating PoPs has to traverse longer paths.

Since all the ISP routers maintain routes to the top 1.5% of popular prefixes that carry 75.5% of the ISP's traffic, 24.5% of the ISP's traffic is impacted due to the use of virtual aggregation. We discuss how the ISP can control this in the next section. However, as shown in figure 9, the traffic stretched is much lower and reduces with increasing percentage of aggregating PoPs. For more than 28% aggregating PoPs, the traffic stretched is less than 4% of the ISP's total traffic. Hence, the deployment of virtual aggregation leads to minimal impact on the traffic distribution across the ISP's inter-PoP links.

Thus, the ISP can use the number of aggregating PoPs as a knob to trade-off stretch imposed on traffic for a reduction in FIB size. We imagine that an ISP, when deploying virtual aggregation, would use a constraint-solving approach to decide on the deployment parameters. For instance, a trivial constraint that an ISP may be interested in is minimising the worst-case FIB size while ensuring that the worst-case stretch is less than a certain value. This may be useful to ensure that its existing SLAs with managed Internet customers are not breached. As a specific example, for the ISP under study, a constraint of 5msec worst-case stretch can be satisfied with the largest 57% of the ISP's PoPs serving as aggregating PoPs resulting in a worst-case FIB size of 9.5% of the DFZ routing table size. Table I shows the parameters for two such solutions.

Another way to quantify the benefits of virtual aggrega-

	Worst-case stretch (msec)	Today	Virtual Aggregation				
			1	4	7	10	20
239K FIB	Quad. Fit Expo. Fit	Expired Expired	2020 2022	2026 2025	2027 2026	2035 2029	2036 2030
1M FIB	Quad. Fit Expo. Fit	2015 2012	2044 2033	2055 2036	2058 2037	2074 2040	2077 2041

TABLE II

ESTIMATES FOR ROUTER LIFE WITH VIRTUAL AGGREGATION

tion is to determine the extension in the life of a router with a specified memory due to the use of virtual aggregation. As proposed in [23], we used data for the DFZ routing table size from Jan'02 to Dec'07 [22] to fit a quadratic model to routing table growth. Further, it has been claimed that the DFZ routing table has seen exponential growth at the rate of 1.3x every two years for the past few years and will continue to do so [32]. We use these models to extrapolate future DFZ routing table size. We consider two router families: Cisco's Cat6500 series with a supervisor 720-3B forwarding engine that can hold upto 239K IPv4 FIB entries and hence, was supposed to be phased out by mid-2007 [5], though some ISPs still continue to use them. We also consider Cisco's current generation of routers with a supervisor 720-3BXL engine that can hold 1M IPv4 FIB entries. For each of these router families, we calculate the year to which they would be able to cope with the growth in the DFZ routing table with the existing setup and with virtual aggregation. Table II shows the results. For virtual aggregation, relaxing the worst-case stretch constraints reduces FIB size and hence, extends the router life. The table shows that if the DFZ routing table were to grow at the aforementioned exponential rate, virtual aggregation can extend the life of the previous generation of routers to 2022 and beyond with a small worst-case stretch and negligible average-case stretch. Of course, the number of factors involved implies that it is very difficult to accurately predict future routing table size and the growth rate could certainly be more than what we have used above. However, note that virtual aggregation only needs to extend router life beyond the point where the routers would need to be updated for other reasons such as newer technologies and higher data rates.

B. Popular Prefixes

Past studies of ISP traffic patterns from as early as 1999 have observed that a small fraction of Internet prefixes carry a large majority of ISP traffic [9,12,37,46]. We used Netflow records collected across the routers of the same tier-1 ISP as in the last section for a period of two months (20th Nov'07 to 20th Jan'07) to generate per-prefix traffic statistics and observed that this pattern continues to the present. The line labeled "Day-based" in figure 10 plots the average fraction of the ISP's traffic destined to a given fraction of popular prefixes when the set of popular prefixes is calculated on a daily basis. The figure shows that 1.5% of most popular prefixes carry 75.5% of the traffic while 5% of the prefixes carry 90.2% of the traffic.

Virtual Aggregation exploits the notion of prefix popularity to reduce its impact on the ISP's traffic. The studies

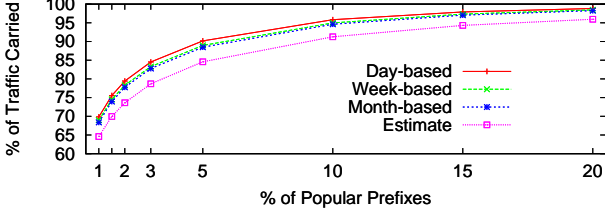


Fig. 10. Popular prefixes carry a large fraction of the ISP's traffic.

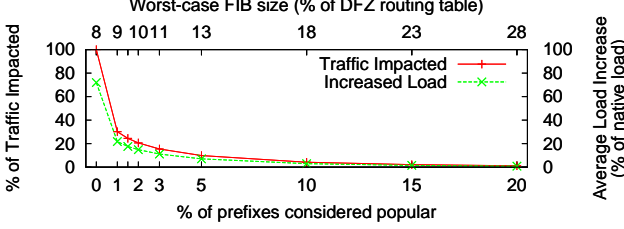


Fig. 11. Traffic impacted and load reduces as routes to more prefixes are maintained by all routers.

mentioned above have shown that prefix popularity is stable enough to be used for traffic engineering purposes. However, when being used for route configuration, it would be preferable to be able to calculate the popular prefixes over a week, month, or even longer durations. To explore the feasibility of such an approach, we calculate the average traffic carried by popular prefixes when the popularity is calculated weekly and monthly. These are plotted in figure 10. We found that the popular prefixes carry almost the same amount of traffic irrespective of whether popularity is measured on a daily, weekly or monthly basis. Further, the line labeled “Estimate” in the figure shows the amount of traffic carried to prefixes that are popular on a given day over the period of the next month, averaged over each day in the first month of our study. As can be seen, the estimate based on prefixes popular on any given day carries just a little less traffic as when the prefix popularity is calculated daily. This suggests that prefix popularity is stable enough for virtual aggregation configuration and the ISP can use the prefixes that are popular on a given day for a month or so. However, we admit that these results are very preliminary and we need to study ISP traffic patterns over a longer period to support the claims made above.

Virtual aggregation causes traffic to take extra hops and hence, increases the load on the ISP's routers. The ISP can increase the number of prefixes that are considered popular to tune the extra load. These popular routes are maintained by all ISP routers and hence, this represents a trade-off between FIB size and router load. Figure 11 illustrates this. As more prefixes are considered popular (along the lower X-axis) and correspondingly the worst-case FIB size increases (along the upper X-axis), both the traffic impacted and the percentage increase in average edge router load reduces. We don't show the load increase across the core routers since they already carry a lot of traffic and hence, the percentage increase in load is very small. Also note that this increase in load is across the internal interfaces of the edge routers where bandwidth is less of a constraint than

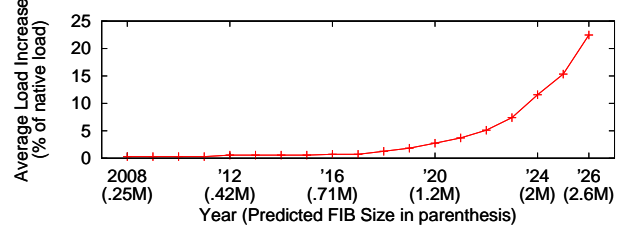


Fig. 12. Extra load due to virtual aggregation increases gradually.

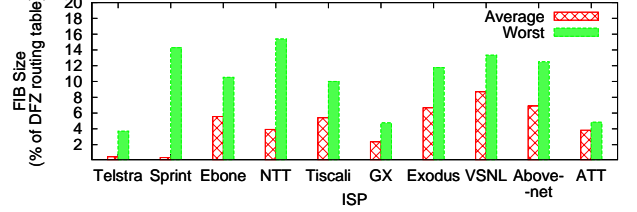


Fig. 13. FIB size for various ISPs using virtual aggregation.

the external interfaces to which external routers connect. However, the extra load may still be a concern for the ISP. For instance, with 5% popular-prefixes, 9.8% of the ISP's traffic is impacted while the average load increase across the edge routers is 7%.

We would like to point out that in a practical deployment, the ISP would not see a sudden step-increase in the router load. The ISP would deploy virtual aggregation such that its routers maintain routes to prefixes according to the virtual aggregation scheme. The rest of the available memory on routers would be used to maintain routes to as many other prefixes as possible. We simulated such a scenario for our ISP assuming that all routers have space for 239K prefixes and the routing table grows at the same exponential rate as mentioned before. Figure 12 shows that the load on the edge routers increases gradually with time. For instance, the load would increase by 1.8% over the next 10 years before increasing more sharply. Note that Internet traffic has been growing at the rate of at least 50-60% per year for the past few years [35] and given that ISPs anyway need to account for this, the extra load imposed by virtual aggregation should not be of major consequence. In other words, virtual aggregation would not impact the hardware upgrade cycle of the ISP to cope with increasing traffic while ensuring that the ISP can still use its existing routers for a long time. Further, the new routers deployed by the ISP don't necessarily need to be equipped with a very large amount of high-end memory and hence, are closer to the inflection point in the cost vs performance curve mentioned in section II.

C. Rocketfuel Study

We studied the topologies of 10 ISPs collected as part of the Rocketfuel project [40] to determine the FIB size savings that virtual aggregation would yield. Note that the fact we don't have traffic matrices for these ISPs implies that we cannot determine the average stretch when an ISP deploys virtual aggregation with a given number of aggregating PoPs. For each ISP, we use the very simple constraint

optimization described in section IV-A to determine how the ISP could deploy virtual aggregation with a given limit on the worst-case stretch. Here we focus on the FIB size of the ISP routers such that the worst-case stretch is less than 5 msecs. Figure 13 shows that the worst-case FIB size is always less than 15% of the DFZ routing table. The worst-case FIB size is relatively higher for NTT and Sprint because they have a global footprint with a few small PoPs outside their main area of influence. For instance, Sprint has a few small PoPs in the Asia-Pacific region and designating one (or, a few) of these as an aggregating PoP to satisfy the stretch constraint limits the reduction in FIB size. However, the Rocketfuel topologies are not complete and are missing routers. Hence, while the results presented here are encouraging, they should be treated as conservative estimates of the savings that virtual aggregation would yield for these ISPs.

D. Discussion

Blades as aggregation points: Routers today tend to have multiple blades with each blade maintaining its own copy of the entire routing table [52]. With virtual aggregation, the routers could be configured to use each of its blades as an aggregation point for different virtual prefixes. This could potentially reduce the FIB size on individual blades by another order of magnitude. However, the lack of information regarding the number of blades supported by the routers in our study’s ISP prevents us presenting specific numbers regarding the router FIB size with such an approach.

Small ISPs: The analysis in the previous section showed that virtual aggregation can significantly reduce FIB size for a few ISPs. Most of these ISPs are large tier-1 and tier-2 ISPs. However, smaller tier-2 and tier-3 ISPs are also part of the Internet DFZ and hence, their routers need to maintain the entire routing table. The fact that these ISPs have small PoPs would seem to suggest that virtual aggregation would not be very beneficial.

However, small ISPs that do have a few PoPs can deploy virtual aggregation by relaxing the constraint that aggregating PoPs have an aggregation point for each virtual prefix. Instead, each PoP could aggregate some of the virtual prefixes. The PoPs of such ISPs are typically geographically close to each other and hence, the stretch imposed by such an approach would be minimal. Actually, the fact that these are not tier-1 ISPs implies they are a customer of at least one other ISP. Hence, the ISP could substantially shrink the FIB size on its routers by applying virtual aggregation to the small number of prefixes advertised by their customers and peers while using default routes for the rest of the prefixes.

V. DEPLOYMENT

To verify the claim that virtual aggregation is a configuration-only solution, we deployed virtual aggregation across two separate testbed networks. The first test network was built on WAIL [1] and comprises of an ISP

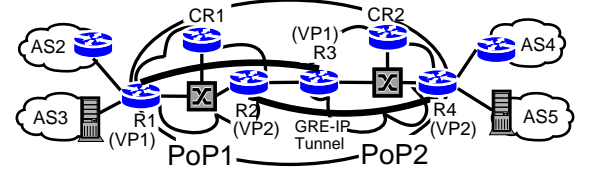


Fig. 14. WAIL topology configured in accordance with virtual aggregation. All routers in the figure are Cisco 7300s. Routers R1 and R3 aggregate virtual prefix VP1 while routers R2 and R4 aggregate VP2.

with two PoPs. The network setup is shown in figure 14. Each PoP has two Cisco 7301 routers and two external routers connected to it. Each PoP also has a conduit-router—we used both Cisco 7301 routers and a Linux PC to serve as conduit-routers. The ISP uses two virtual prefixes: 0.0.0.0/1 (VP1) and 128.0.0.0/1 (VP2) with one router in each PoP serving as an aggregation point for each virtual prefix. The ISP’s edge routers (R1 and R4) are configured to “route” packets at layer-3 and “switch” packets at layer-2. Further, the PoP-facing interface of external routers is on the same subnet as the PoP’s internal routers and hence, there is layer-2 connectivity between the PoP’s internal and external routers. The external routers exchange routes through an eBGP peering with the conduit-router of the PoP they are connected to. The internal routing of the ISP is set up according to the description in section III with each conduit-router serving as a route-reflector for the PoP’s internal routers, iBGP peerings between the conduit-routers themselves and GRE-IP tunnels between routers that are aggregation points for the same virtual prefix.

The conduit-routers use BGP *route-maps* on their peerings to ensure that routes are distributed according to the virtual aggregation scheme, including the distribution of popular prefixes to all routers. For instance, router R1 is an aggregation point for virtual prefix VP1 and hence, should only be forwarded routes for popular prefixes, routes for prefixes in VP1 and a route for the virtual prefix VP2 but not any prefix inside VP2. The conduit-router CR1 satisfies the restrictions regarding virtual prefix VP2 using outbound route-map “filter-vp2” on its peering with R1:

```
! first half of VP2
access-list vp2-1 permit 128.0.0.0/2
! second half of VP2
access-list vp1-2 permit 192.0.0.0/2

! block sub-prefixes of VP2
route-map filter-vp2 deny 10
match ip address vp2-1
route-map filter-vp2 deny 20
match ip address vp2-2
! advertise VP2 itself
route-map filter-vp2 permit 30
```

Given the design presented in this paper, the number of distinct route-map rules used on a conduit-router is $(5V + p + E)$, where V is the number of virtual prefixes, p is the number of popular prefixes and E is the number of external routers connected to the PoP. Note that these route-maps don’t need to be on fast memory and even then, the memory overhead due to this is very small ($<1\text{MB}$). However, there

are lots of other configuration details such as changing the next-hop attribute of the routes at the conduit-routers, the configuration of non-aggregating PoPs, etc. that we don't describe here. [53] describes such details and shows the configuration scripts used.

Restrictions on the number of available routers meant that the topology above is very small. We also configured virtual aggregation across a network comprising of Linux PCs running kernel 2.6.20 with the `quagga` BGP software router. The ISP here had four PoPs with one of them configured as a non-aggregating PoP. Each PoP of the ISP had at least four routers (two core and two edge routers) and four external routers connected to it. Aside from the differences in syntax, the configuration on Linux software routers used the same standard BGP features as the configuration on the Cisco routers.

The virtual aggregation deployment on both testbeds involves manual configuration of GRE-IP tunnels between the aggregation points in various PoPs. This represents a significant configuration overhead and a robustness concern for the ISP. However, such manual configuration can be avoided by using BGP-MPLS for the virtual aggregation control plane. To illustrate this, we configured virtual aggregation on the WAIL topology using BGP-MPLS. The key idea here is to treat each virtual prefix like a VPN and hence, assign it a VPN label (i.e., a VRF number). The ISP has the same setup involving per-PoP conduit-routers as described in this paper with the conduit-routers ensuring that individual routers only receive routes for the virtual prefix they are aggregating. However, MP-BGP peerings are used to distribute the routes internally. This, in turn, ensures that the configuration of the actual MPLS forwarding plane is done by LDP and hence, MPLS tunnels are automatically established between the appropriate aggregation points.

VI. DISCUSSION AND FUTURE WORK

Pros. Virtual Aggregation can be *incrementally deployed* by an ISP since it does not require the cooperation of other ISPs and router vendors. The ISP does not even need to change the structure of its PoPs or its topology. What's more, an ISP could experiment with virtual aggregation on a limited scale (a few virtual prefixes or a limited number of PoPs) to gain experience and comfort before expanding its deployment. None of the attributes in the BGP routes advertised by the ISP to its neighbors are changed due to the adoption of virtual aggregation. The routes chosen by the ISP for each prefix might not be the same as the ones chosen if the ISP had a vanilla BGP deployment involving an iBGP mesh between the ISP's routers. However, the same is true when the ISP uses route reflectors and other mechanisms for scalable route distribution. Also, the use of virtual aggregation by the ISP does not restrict its routing policies and route selection. As a matter of fact, the conduit-router of each PoP provides a convenient enforcement point for the ISP's policies. Finally, there is *incentive for deployment*

since the ISP improves its own capability to deal with routing table growth.

Concerns. The use of virtual aggregation does impose significant configuration burden on the ISP. This includes configuring route filters at the conduit-routers for appropriate control-plane operation and in case of a non BGP-MPLS based deployment, configuring tunnels between the routers for appropriate data-plane operation. Further, the ISP needs to make a number of deployment decisions such as choosing the virtual prefixes to use, deciding where to keep aggregation points for each virtual prefix, which prefixes to consider popular, and so on. Thus, the ISP would need a network management system that can take various constraints such as stretch and load constraints and other high-level goals to generate the required router configurations. Apart from such one-time or infrequent decisions, virtual aggregation may also influence very important aspects of the ISP's day-to-day operation such as maintenance, debugging, etc.

Of course, there is a cost associated with all this. However, this cost may be significantly lower than the cost of upgrading routers which, apart from the capital costs, requires reconfiguring every customer on every router twice. Hence, virtual aggregation presents a cost trade-off between the increased management costs and the decreased cost of router upgrades and we intend to consult ISP network managers regarding our conjecture that this is a beneficial trade-off.

Another important concern arising out of the use of virtual aggregation is the tunneling overhead. However, the extensive use of tunnels (MPLS, GRE-IP, IPsec, VLAN tunneling) in ISP networks has meant that most current generation routers are already equipped with interfaces that have extensive tunneling and detunneling capabilities at line rates [13].

In terms of technical metrics, virtual aggregation represents a trade-off between FIB size reduction on one hand and increased router load and traffic stretch on the other. The fact that Internet traffic follows a power-law distribution makes this a very beneficial trade-off. If Internet traffic were uniformly distributed across all prefixes, the stretch imposed on traffic would still be reasonable (average stretch of 0.08 msec for the ISP studied in section IV-A) but the same cannot be said for the load increase on routers. However, the power-law observation has held up in measurement studies from 1999 [9] to 2007 (in this paper) and hence, Internet traffic has followed this distribution for at least the past eight years in spite of the rise in popularity of P2P and video streaming. We believe that, more likely than not, future Internet traffic will be power-law distributed and hence, virtual aggregation will represent a good trade-off for ISPs.

The discussion in section III-G and the results presented in section IV assume that prefixes are uniformly distributed across the virtual prefixes. This is not the case. For instance, given the prefixes present in the DFZ routing table today, we calculated that using /7s as virtual prefixes would imply

a significant variation in the number of prefixes that each virtual prefix contains. However, all the virtual prefixes need not be of the same length. The only constraint is that they cover the IPv4 address space and all virtual prefixes should be less specific than actual prefixes. Hence, the ISP can ensure a relatively uniform distribution of prefixes across the virtual prefixes through smarter virtual prefix allocation.

Other design points. The virtual aggregation architecture presented in this paper represents one point in the design space that we focussed on for the sake of concreteness. The basic idea of dividing the routing table such that individual routers only need to maintain part of the routes can be achieved using a few alternative approaches. For instance, we are working on a proposal that assumes cooperation amongst ISPs. Apart from reducing the routing table burden on routers even more, this can vastly improve the convergence properties of the routing system. Similarly, there are designs that involve changes to routers, changes to the ISP topology and so on. Below we very briefly describe a couple other approaches. We intend to study the merits and demerits of such alternative designs in future work.

- *ISPs with internal layer-2 connectivity.* The presented design requires layer-2 connectivity between a PoP’s internal and external routers. Many ISPs today are moving towards internal layer-2 connectivity between all their PoPs and hence, our design could be extended to ensure that all the ISP’s routers can reach directly connected external routers at layer-2. This would do away with the need for non-aggregating PoPs, thereby reducing stretch, doing away with the need to tunnel packets between routers and greatly simplifying virtual aggregation configuration. It would also provide the ISP with more flexibility regarding the placement of aggregation points for different virtual prefixes.

- *Adding routers.* An ISP can avoid the stretch and additional complexity resulting from the use of non-aggregating PoPs by adding routers to small PoPs. Actually, the ISP could even ensure that these are not “first-class” routers. Instead it could use “slow-fat” routers that are only responsible for routing packets to a fraction of the unpopular prefixes. This would ensure that the routers don’t have high performance requirements. Consequently, these need not be expensive hardware routers; the ISP could make do with a stack of inexpensive software routers [17] and hence, achieve scalability through stackability.

VPN Scalability. Another major problem for ISPs with regards to routing scalability is the need to maintain VPN routing tables for their VPN customers. The use of BGP-MPLS [8] for VPNs ensures that only Provider-Equipment (PE) routers directly connected to the VPN’ed customers need to keep VPN routes. However, these VPN tables are typically several times larger than the global routing table and hence, the scaling problem. The trick of using a conduit-router and layer-2 switching can be exploited to cause VPN traffic to travel at layer-2 from CE (customer-equipment) to

CE while ensuring that a given customer site only peers with the VPN provider. This eliminates the FIB in the PE equipment altogether. We intend to validate this approach for a couple of tier-1 ISPs that are also major VPN providers as part of future work.

VII. RELATED WORK

Over the years, several articles have documented the existing state of inter-domain routing and delineated requirements for the future [2,4,11,15,30,33]. A number of efforts have tried to directly tackle the routing scalability problem. Section I mentioned some of these proposals [6,7,10,14,19,31,34]. Our work resembles some aspects of CRIO [50] which uses virtual prefixes and tunneling to decouple network topology from addressing. However, CRIO requires adoption by all provider networks. Also, like [6,10,31,34], it requires a separate new mapping service to determine tunnel endpoints. APT [24] presents such a mapping service. Our proposal avoids the need for a separate service and effectively achieves the mapping through existing control-plane mechanisms. Similar to CRIO, Verkaik et. al. [47] group prefixes with similar behavior into policy atoms and use these atoms and tunneling of packets to reduce routing table size.

More generally, the use of tunnels has long been proposed as a routing scaling mechanism. As mentioned in the previous section, VPN technologies such as BGP-MPLS VPNs [8] use tunnels to ensure that only PE routers need to keep the VPN routes. As a matter of fact, ISPs can and probably do use tunneling protocols such as MPLS and RSVP-TE to engineer a BGP-free core [39]. However, edge routers still need to keep the full RIB and FIB. With virtual aggregation, none of the routers on the data-path need to maintain the full RIB and FIB.

Some router vendors use FIB compression to reduce the FIB size on routers [39]. This avoids installation of redundant more specific prefixes in the FIB. Another technique to deal with routing table growth is to cache routes to popular prefixes in expensive and fast memory such as SRAM or TCAM while the entire FIB is maintained on cheaper DRAMs [39]. An interesting set of approaches that trade-off stretch for routing table size are *Compact Routing* algorithms. The key idea behind such algorithms is the adoption of a more flexible notion of best path. Krioukov et. al. [26] analyze the performance of such an algorithm for Internet-like graphs; see [27] for a survey of the area.

While scalability might be the most important problem afflicting inter-domain routing, several other aspects of routing have also received a lot of attention. For instance, proposals for improving BGP convergence time [41,43], enabling host control over routing [49], and improving routing security [25,42] represent a few examples. RCP [3] and 4D [18] argue for logical centralization of routing in ISPs to provide scalable internal route distribution and a simplified control plane respectively. We note that virtual aggregation fits well into these alternative routing models.

VIII. CONCLUSIONS

This paper presents a simple approach that can be used by ISPs to cope with increasing routing table size. While it is often (implicitly) assumed that routers in the default-free zone of the Internet need to maintain routes to all prefixes being advertised into the Internet, we show that an ISP can modify its internal routing such that individual routers only need to maintain a part of the global routing table. Apart from requiring configuration changes only, the design presented in this paper allows ISPs to experiment with virtual aggregation on a limited scale. As a matter of fact, we plan to utilise this flexibility as part of our efforts to deploy virtual aggregation on an operational network and are in discussions with Internet2 and other ISP operators regarding this. Such a deployment would go a long way in concretely answering questions regarding the impact of virtual aggregation on the ISP's operation, especially the management and robustness consequences of the increased configuration. However, these questions notwithstanding, we believe that the simplicity of the proposal and its possible short-term impact on routing scalability suggest that it is an alternative worth considering.

REFERENCES

- [1] BARFORD, P. Wisconsin Advanced Internet Laboratory (WAIL), Dec 2007. <http://wail.cs.wisc.edu/>.
- [2] BONAVENTURE, O., QUOTIN, B., AND UHLIG, S. Beyond Inter-domain Reachability. In *Proc. of Workshop on Internet Routing Evolution and Design (WIRED)* (2003).
- [3] CAESAR, M., CALDWELL, D., FEAMSTER, N., REXFORD, J., SHAIKH, A., AND VAN DER MERWE, J. Design and Implementation of a Routing Control Platform. In *Proc. of Symp. on Networked Systems Design and Implementation (NSDI)* (2005).
- [4] DAVIES, E., AND DORIA, A. Analysis of Inter-Domain Routing Requirements and History. Internet Draft draft-irtf-routing-history-07.txt, Jan 2008.
- [5] DE SILVA, S. 6500 FIB Forwarding Capacities. NANOG 39 meeting, 2007. <http://www.nanog.org/mtg-0702/presentations/fib-desilva.pdf>.
- [6] DEERING, S. The Map & Encap Scheme for scalable IPv4 routing with portable site prefixes, March 1996. <http://www.cs.ucla.edu/~lixia/map-n-encap.pdf>.
- [7] DEERING, S., AND HINDEN, R. IPv6 Metro Addressing. Internet Draft draft-deering-ipv6-metro-addr-00.txt, Mar 1996.
- [8] E. ROSEN AND Y. REKHTER. RFC 2547 - BGP/MPLS VPNs, Mar 1999.
- [9] FANG, W., AND PETERSON, L. Inter-As traffic patterns and their implications. In *Proc. of Global Internet* (1999).
- [10] FARINACCI, D., FULLER, V., ORAN, D., AND MEYER, D. Locator/ID Separation Protocol (LISP). Internet Draft draft-farinacci-lisp-02.txt, July 2007.
- [11] FEAMSTER, N., BALAKRISHNAN, H., AND REXFORD, J. Some Foundational Problems in Interdomain Routing. In *Proc. of Workshop on Hot Topics in Networks (HotNets-III)* (2004).
- [12] FELDMANN, A., GREENBERG, A., LUND, C., REINGOLD, N., REXFORD, J., AND TRUE, F. Deriving traffic demands for operational IP networks: methodology and experience. *IEEE/ACM Trans. Netw.* 9, 3 (2001).
- [13] FRANCIS, P., AND BONAVENTURE, O. An evaluation of IP-based Fast Reroute Techniques. In *Proc. of CoNEXT* (2005).
- [14] FRANCIS, P. Comparison of geographical and provier-rooted Internet addressing. *Computer Networks and ISDN Systems* 27, 3 (1994).
- [15] G. HUSTON. RFC 3221 - Commentary on Inter-Domain Routing in the Internet, Dec 2001.
- [16] GHAZI, A. Best Practices for ISPs. APNIC 14 meeting, 2002. <http://www.apnic.net/meetings/14/sigs/routing/>.
- [17] GILLIAN, B. VYATTA: Linux IP Routers, Dec 2007. http://freedomhpc.pbwiki.com/f/linux_ip_routers.pdf.
- [18] GREENBERG, A., HJALMTYSSON, G., MALTZ, D. A., MEYERS, A., REXFORD, J., XIE, G., YAN, H., ZHAN, J., AND ZHANG, H. A clean slate 4D approach to network control and management. *ACM SIGCOMM Computer Communications Review* (October 2005).
- [19] HAIN, T. An IPv6 Provider-Independent Global Unicast Address Format. Internet Draft draft-hain-ipv6-PI-addr-02.txt, Sep 2002.
- [20] HUGHES, D., Dec 2004. PACNOG list posting <http://mailman.apnic.net/mailling-lists/pacnog/archive/2004/12/msg00000.html>.
- [21] HUSTON, G. ISP Column: Whither Routing, Nov 2006. <http://www.potaroo.net/ispcol/2006-11/raw.html>.
- [22] HUSTON, G. BGP Reports, Dec 2007. <http://bgp.potaroo.net/>.
- [23] HUSTON, G., AND ARMITAGE, G. Projecting Future IPv4 Router Requirements from Trends in Dynamic BGP Behaviour. In *Proc. of ATNAC* (2006).
- [24] JEN, D., MEISEL, M., MASSEY, D., WANG, L., ZHANG, B., AND ZHANG, L. APT: A Practical Transit Mapping Service. Internet Draft draft-jen-apt-01.txt, Nov 2007.
- [25] KENT, S., LYNN, C., AND SEO, K. Secure border gateway protocol (S-BGP). *IEEE Journal on Selected Areas in Communication* 18, 4 (2000).
- [26] KRIOUKOV, D., FALL, K., AND YANG, X. Compact routing on Internet-like graphs. In *Proc. of IEEE INFOCOM* (2004).
- [27] KRIOUKOV, D., AND KC CLAFFY. Toward Compact Interdomain Routing, Aug 2005. <http://arxiv.org/abs/cs/0508021>.
- [28] LANDLER, P. DRAM Productivity and Capacity/Demand Model. In *Proc. of Global Economic Workshop* (1999).
- [29] LI, T. Router Scalability and Moore's Law, Oct 2006. http://www.iab.org/about/workshops/routingandaddressing/Router_Scalability.pdf.
- [30] MAO, Z. M. Routing Research Issues. In *Proc. of WIRED* (2003).
- [31] MASSEY, D., WANG, L., ZHANG, B., AND ZHANG, L. A Proposal for Scalable Internet Routing & Addressing. Internet Draft draft-wang-ietf-efit-00, Feb 2007.
- [32] MEYER, D., ZHANG, L., AND FALL, K. Report from the IAB Workshop on Routing and Addressing. Internet Draft draft-iab-raws-report-02.txt, Apr 2007.
- [33] NARTEN, T. Routing and Addressing Problem Statement. Internet Draft draft-narten-radir-problem-statement-01.txt, Oct 2007.
- [34] O'DELL, M. GSE-An Alternate Addressing Architecture for IPv6. Internet Draft draft-ietf-ipngwg-gseaddr-00.txt, Feb 1997.
- [35] ODLYZKO, A. Minnesota Internet Traffic Studies (MINTS), Dec 2007. <http://www.dtc.umn.edu/mints>.
- [36] P. TRAINA. RFC 1965 - Autonomous System Confederations for BGP, Jun 1996.
- [37] REXFORD, J., WANG, J., XIAO, Z., AND ZHANG, Y. BGP routing stability of popular destinations. In *Proc. of Internet Measurement Workshop* (2002).
- [38] ROBINSON, S. Manning Up for 10 Gigabit Ethernet, Dec 2007. <http://www.commsdesign.com/showArticle.jhtml?articleID=16502737>.
- [39] SCUDDER, J. Router Scaling Trends. APRICOT Meeting, 2007. http://submission.apricot.net/chat07/slides/future_of_routing.
- [40] SPRING, N., MAHAJAN, R., AND WETHERALL, D. Measuring ISP topologies with Rocketfuel. In *Proc. of ACM SIGCOMM* (2002).
- [41] SUBRAMANIAN, L., CAESAR, M., EE, C. T., HANDLEY, M., MAO, M., SHENKER, S., AND STOICA, I. HLP: A Next Generation Inter-domain Routing Protocol. In *Proc. of ACM SIGCOMM* (2005).
- [42] SUBRAMANIAN, L., ROTH, V., STOICA, I., SHENKER, S., AND KATZ, R. Listen and whisper: Security mechanisms for BGP. In *Proc. of USENIX/ACM NSDI* (2004).
- [43] SUN, W., MAO, Z. M., AND SHIN, K. Differentiated BGP Update Processing for Improved Routing Convergence. In *Proc. of ICNP* (2006).
- [44] SYSTEMS, C. BGP Multihoming. PACNOG 1 meeting, 2005. <http://www.pacnog.net/pacnog1/day5/b4-6up.pdf>.
- [45] T. BATES AND R. CHANDRA AND E. CHEN. RFC 2796 - BGP Route Reflection - An Alternative to Full Mesh IBGP, Apr 2000.
- [46] TAFT, N., BHATTACHARYA, S., JECHEVA, J., AND DIOT, C. Understanding traffic dynamics at a backbone PoP. In *Proc. of Scalability and Traffic Control and IP Networks SPIE ITCOM* (2001).
- [47] VERKAIK, P., BROID, A., KC CLAFFY, GAO, R., HYUN, Y., AND VAN DER POL, R. Beyond CIDR Aggregation. Tech. Rep. TR-2004-1, CAIDA, 2004.
- [48] Y. REKHTER AND T. LI AND S. HARES, ED. RFC 4271 - A Border Gateway Protocol 4 (BGP-4), Jan 2006.
- [49] YANG, X. NIRA: a new Internet routing architecture. In *Proc. of the ACM SIGCOMM workshop on Future directions in network architecture (FDNA)* (2003).
- [50] ZHANG, X., FRANCIS, P., WANG, J., AND YOSHIDA, K. Scaling Global IP Routing with the Core Router-Integrated Overlay. In *Proc. of ICNP* (2006).
- [51] Cisco BGP Documentation, Dec 2007. <http://www.cisco.com/univercd/cc/td/doc/cisintwk/ito.doc/bgp.htm#wp1020610>.
- [52] Network Processor Blades, Dec 2007. http://dnd.ecitele.com/products/literature/eci_npb_final.pdf.
- [53] ANONYMIZED. Routing Scalability through Virtual Aggregation. Tech. rep., Available upon request, 2007.